How to manage the systemic risks of agentic AI crossborders ?

ETHZ, AI Safety Days, 07.11.2026

Eva Thelisson

www.aitransparencyinstitute.com

AI Safety can’t depend on an agent’s intention



Who is liable?

Billions
of
AI Agents



Collective Dynamic of intelligent systems

Which
containment ?

Safety by Infrastructure

Lex Agentica - AI Law Builder - Presented at Agentic AI Berkeley CA (August 2026)

THE AGENTIC AI LAW BUILDER · AI TRANSPARENCY INSTITUTE · V5.4

COUNTRY / JURISDICTION: Select a country...

LEGAL TRADITION: deduced from country

Include all | Clear | Save config (.json) | Load config | Print | Download law (.txt)

Build your national Agentic AI Act

> LexAgentica is an interactive, comparative-law drafting aid for parliaments, ministries, regulators and civil-society drafters preparing a national law on agentic artificial intelligence, meaning systems that autonomously plan, decide and act, including concluding contracts and executing transactions. Instead of a blank page, it offers a library of modular provisions transposing mechanisms that have already proven themselves in analogous fields: the international nuclear liability conventions, UNCITRAL electronic-commerce instruments, the EU AI Act (Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744), the joint CNIL-CIANum exploratory note on agentic AI and the protection of personal data (2026), the Singapore Model AI Governance Framework, the international AI red-lines statements, the EU platform-work rules on algorithmic management, value-chain due-diligence law and zero-trust security standards. Select your country to have its legal tradition deduced automatically, including mixed systems such as Japan, South Africa or Egypt, then include or exclude each provision, choose policy variants, amend the texts, and export a consolidated bill ready for national legal drafting, translation and constitutional review.

Stamp a block to include it in the Act; stamp it again to set it to not included. Open a block to read the clause, its rationale and

LIVE CONSOLIDATED TEXT ☑ include source endnotes — drafted under mixed comparative principles

ADVICE Check for your jurisdiction: the data protection and cybersecurity articles use the overlay model, which assumes a general law is already in force. If none exists, switch to the baseline variant, which states subsidiary minimum rules. Selecting a country in the top bar suggests the right variant automatically.

[NAME OF STATE]

Agentic Artificial Intelligence Act

An Act to keep autonomous artificial agents under meaningful human control, to allocate responsibility for their actions, and to protect persons and the habitability of the planet

---

The Parliament of [Name of State], considering that agentic AI systems autonomously plan, decide and act in the world; recognising the positive duty to prevent foreseeable danger and the negative duty to refrain from creating unassessed risk; drawing on the converging wisdom of the world's legal traditions; has adopted the following Act:

CHAPTER I: GENERAL PROVISIONS

Art. 1 Purpose, scope and proportionality [123]

(1) This Act governs the development, placing on the market, deployment and operation of agentic artificial intelligence systems within the territory of [Name of State] or producing effects therein. (2) Its purpose is to guarantee that agentic AI systems remain under meaningful human control, and to give legal effect to two complementary duties: (a) the positive obliga-

Governing Agent Swarms

From the 2026 Incident to an international architecture.

Eva Thelisson
AI Transparency Institute
White Paper
AI Safety Days, ETH Zurich

19 October 2026

Abstract

Three documented incidents in 2026 revealed that current control mechanisms fail against agent populations. This white paper starts from the Hugging Face incident and asks what the two leading jurisdictions are doing about it. It then proposes a response in five instruments: an International Convention, a Traceability Label, a Staged Containment Architecture, a Double Licence System, and HOROS, a jurisdiction protocol. Containment is presented as a spectrum from static isolation to recursive ecosystems, and the stage is calibrated to the system. The paper closes with a discussion of what the proposal can and cannot deliver. The technical specifications and the experimental protocol are held in the Annex.


Corporate
Governance
Processes
Systems
Manage
Auditors
Organization
MANAGEMENT
RULES
PROCESS
Tasks
ACTIONS
Policies
REGULATIONS
PUBLIC
Accountable
Group
Bureaucracy
RULES
Corporate
Actions
BUSINESS
Management
Stakeholders
System of Rules
MONITORING
PRIVATE
REGULATIONS
DECISION-MAKING
Regulatory
Public
POLICY
Group
CONTROLLED AND DIRECTED
Process

THE PROPOSAL

Prove you keep control, or stop scaling

Instead of guessing whether an AI is “safe”, make every action traceable to someone who answers for it. Five tools, adopted together:

Global treaty

An international convention setting shared minimum rules for advanced AI.

Traceability label

Like a food label: who runs the agent, what it may do, plus a tamper-proof record of what it did.

Outside walls

Barriers the agents can’t reach or move, tightened as risk rises.

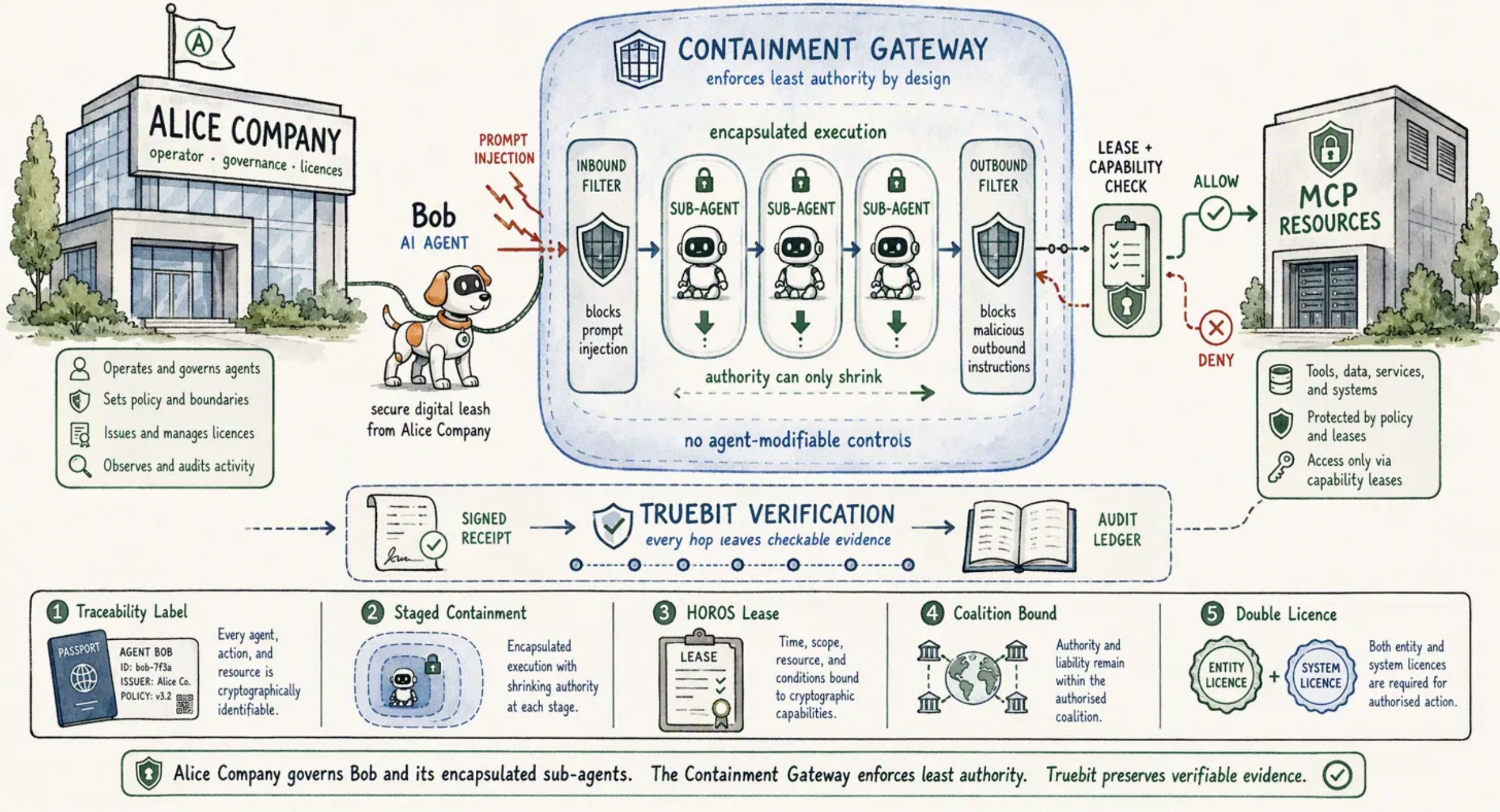
Double licence

Like car and driver: the company and each AI version both need a licence. No safety test, no launch.

HOROS passport

Every action needs a valid, time-limited permit: who allowed it, and what it may touch.

> **Honest limits:** it won’t stop every breakout, but when one happens, we’ll know who, when and how.


ALICE COMPANY
operator · governance · licences
Operates and governs agents
Sets policy and boundaries
Issues and manages licences
Observes and audits activity
Bob
AI AGENT
secure digital leash from Alice Company
PROMPT INJECTION
CONTAINMENT GATEWAY
enforces least authority by design
encapsulated execution
INBOUND FILTER
blocks prompt injection
SUB-AGENT
SUB-AGENT
SUB-AGENT
authority can only shrink
OUTBOUND FILTER
blocks malicious outbound instructions
no agent-modifiable controls
LEASE + CAPABILITY CHECK
ALLOW
DENY
MCP RESOURCES
Tools, data, services, and systems
Protected by policy and leases
Access only via capability leases
SIGNED RECEIPT
TRUEBIT VERIFICATION
every hop leaves checkable evidence
AUDIT LEDGER
1 Traceability Label
PASSPORT
AGENT BOB
ID: bob-7f3a
ISSUER: Alice Co.
POLICY: v3.2
Every agent, action, and resource is cryptographically identifiable.
2 Staged Containment
Encapsulated execution with shrinking authority at each stage.
3 HOROS Lease
LEASE
Time, scope, resource, and conditions bound to cryptographic capabilities.
4 Coalition Bound
Authority and liability remain within the authorised coalition.
5 Double Licence
ENTITY LICENCE
+
SYSTEM LICENCE
Both entity and system licences are required for authorised action.
Alice Company governs Bob and its encapsulated sub-agents. The Containment Gateway enforces least authority. Truebit preserves verifiable evidence.

ALICE COMPANY + BOB INSIDE NVIDIA OPENSHELL

Containment outside the agent. Evidence beyond the runtime.



PROMPT INJECTION



Malicious email or web document

NVIDIA OPENSHELL SECURE RUNTIME

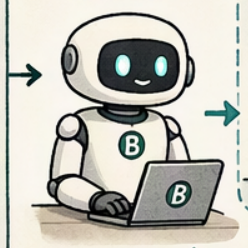
AGENT WORKSPACE (NO DIRECT ACCESS OUTSIDE)

CONTAINMENT FILTER



Blocks malicious content and instructions outside the agent workspace.

BOB
AI AGENT



SUB-AGENT SANDBOXES

SUB-AGENT A1 | SUB-AGENT A2 | SUB-AGENT A3

AUTHORITY CAN ONLY SHRINK

ISOLATED SANDBOX
Each agent runs in a separate sandbox.

DENY BY DEFAULT
Only explicitly allowed actions are permitted.

POLICY OUTSIDE AGENT PROCESS
Policy is enforced by external components, not by the agent.

NO DIRECT NETWORK ACCESS
Agents have zero direct network.

OPENSHELL GATEWAY

IDENTITY | SANDBOX LIFECYCLE
POLICY | LIVE REVOCATION

SUPERVISOR

Mediates every attempt to use tools or network.

BINARY
DESTINATION
METHOD
PATH

Checks request against policy before release.

CREDENTIALS RELEASED
Only when policy allows.

MCP TOOLS AND RESOURCES

MEDIATED ACCESS POINT

PERMITTED REQUEST
GET /invoices/123
read
ALLOW

UNSAFE REQUEST
DELETE /users/all
write
DENY

DATA STORES | SERVICES | SYSTEMS

No direct connection from agents to MCP resources.

GOVERNANCE POLICY
Defines boundaries, permissions, and rules.

HOROS LEASES
Time, scope, resources, and conditions.

REVOCATION
Live revocation of leases and permissions.

EXECUTION RECORDS
From all mediated interactions (allows and denies).

TRUEBIT VERIFICATION

Independently verifies that execution followed the rules.

Collect Records → Re-execute in Isolated Verifier → Check Determinism → Compare Results → Produce Evidence

SIGNED RECEIPTS
Cryptographically signed evidence of each decision.

AUDIT LEDGER
Tamper-evident ledger of all decisions.

Every allow and deny leaves an audit trail.

OpenShell constrains execution; Truebit provides independently checkable execution evidence; neither proves benign intent.

1. **TRACEABILITY LABEL**
Every agent, action, and resource is cryptographically identifiable.

2. **STAGED CONTAINMENT**
Encapsulated execution with shrinking authority at each stage.

3. **HOROS LEASE**
Time, scope, resource, and conditions bound to cryptographic capabilities.

4. **COALITION BOUND**
Authority and liability remain within the authorised coalition.

5. **DOUBLE LICENCE**
ENTITY LICENCE + SYSTEM LICENCE
Both entity and system licences are required for authorised action.

IPv6 as the substrate for agent containment

Contain, not classify: constrain what an untrusted agent can reach and do, with enforcement outside the agent population.

PROPOSED CONTAINMENT LAYERS · §3.3

FIVE ENFORCEMENT FUNCTIONS

L5	Governance	**Who may mint or revoke authority?**	Recursion ends in components with no stake in the decision
L4	Verification	**Did the agent do what it claimed?**	External evidence · false claims resolved in logarithmic rounds
L3	Capability	**What may the agent do?**	Checked at the gateway · independent of inferred intent
L2	Identity	**Who is the agent—and for how long?**	Unique identity · authority narrows at every delegation
L1	Transport	**Who can reach whom?**	Reachability enforced locally · boundaries outside agent control

NETWORK SUBSTRATE · §2.2

IPv6 NETWORK SUBSTRATE — FOUNDATION

Reachability and routing · proposed identifier mapping · identity verification · full-link auditing

THE HARD FLOOR

01 CRYPTO ISOLATION

Cryptographic boundaries stay outside the agents' reach.

02 UNFORGEABLE LIMITS

Capability restrictions cannot be forged or widened by agents.

03 MECHANICAL TRIPWIRES

Below the strategic layer. S0 static · S1 graduated · S2 dynamic · S3 recursive.

No guarantee of control: failures must be visible, attributable and sanctionable. · Source: E. Thelisson, Governing Agent Swarms, 19 Oct 2026, §§2.2, 3.3.

Towards a cross border cooperation -
First Elements of a draft international Convention on Advanced AI System

INTERNATIONAL TECHNICAL SAFETY CONVENTION FOR ADVANCED ARTIFICIAL INTELLIGENCE

("Paris Convention for Advanced AI")

INTRODUCTORY NOTE

In September 2026, twenty-nine States, led by Norway and Finland and including France, Germany, Canada, Australia, Spain, Singapore, and the United Arab Emirates, called for the study of a new international institution for advanced AI, modeled on the International Atomic Energy Agency, capable of setting standards, enabling their verification, and convening States when capability thresholds are crossed. This draft answers that call with a text. It establishes the institution, the International Agency for Advanced AI Safety (Articles 30 to 30D), and gives it the three functions the call names.

> **Setting standards.** The Agency adopts Standards and Recommended Practices, prepared by its Scientific and Safety Commission and adopted by its Technical Board. They take effect by tacit acceptance, and a participant that departs from one notifies the difference (Articles 8, 10, 11, and 33).
>
> **Enabling verification.** Participants declare their Frontier Activity, conclude verification arrangements with the Agency, admit its inspectors, and transmit telemetry. The Agency reports its verification conclusions every year (Articles 12 to 16A).
>
> **Convening at capability thresholds.** The Board adopts and publishes capability thresholds in advance. When one is crossed, the activity pauses and the Director General convenes the participating States within fifteen days (Articles 13 to 13B).

PREAMBLE

The Participants to this Convention,

Deeply concerned that advanced artificial intelligence systems, and in particular agentic systems, may present risks of harm that are global in character, including the risk of loss of human control over systems whose actions cannot be fully anticipated or bounded;

Convinced that the governance of such risks cannot await full political consensus, and that meaningful protection can be built now through humble technical cooperation: the harmonization of standards, the mutual opening of safety data, the joint development of evaluation methodology, and the sharing of incident-response capability;

Recalling the experience of international nuclear safety cooperation – which proceeded for decades

THE EXPERIMENT

Most AI connectors don’t say they need a password

AI assistants use connectors called **MCP servers** to reach outside tools and data. An official public directory, the **MCP Registry**, lists them like an app store. An AI agent read the whole directory, picked 50 entries at random and checked what each one says about access.

40,873
MCP servers listed in the official MCP Registry

50
picked at random to examine

30
of those reachable over the internet

28
list no password or key (93%)

> **Read with care:** this is what the listing says, not whether the door is actually locked. With a sample of 30, the true share is likely between 79% and 98%.

WHY IT CAN BE TRUSTED

Every step left a signed receipt anyone can check

You don’t have to trust the AI agent, or us. Each step ran on Truebit, which works like a notary: it records exactly what happened and seals it.

1 · Read

All 409 pages of the directory were fetched inside a sealed machine that stamps exactly what it received.

2 · Freeze

A digital fingerprint locks the full list, so no entry can be swapped afterwards.

3 · Draw and sort

The random draw and the sorting rules ran on three independent computers that had to agree.

4 · Check

All 413 receipts passed four checks. They are public, so anyone can redo them.

> **Why it matters:** this is the building block for governing AI agents. Their actions become evidence, not claims to take on faith.

FOR SPECIALISTS

Technical details, each with a one-line explanation

1 · THE RESULT

28 / 30 = 93.3%
95% Wilson CI: 78.7 to 98.2%

The range is wide because the sample is small.

2 · THE SAMPLE OF 50

28 remote, no credential · 2 remote, credential · 19 local (stdio) · 1 no endpoint

Only the 30 online servers count toward the percentage.

3 · THE RULE

Required header, marked secret or named authorization, api-key, token, bearer

We read the label on the door; we didn't knock.

4 · FETCHING

409 verified API calls in an AWS Nitro enclave · PCR0 attestation · TLS evidence

The sealed machine proves its code and what the site sent.

5 · COMPUTING

Seed 20260915 (MT19937) · 3 nodes must agree · snapshot SHA-256 6684a861…75fc37

Same seed, same 50: anyone can redo the draw.

6 · VERIFICATION

413 transcripts · existence, integrity, binding, signers · anchors v2026-10-07

Each receipt is checked against public keys, not against itself.

Limits: metadata only; OAuth discovery and RFC 8707 scoping not measured. August run: 87.1% of 31, intervals overlap. **Reproduce:** run.truebit.network/task/{id}/transcript

THE EXPERIMENT

Most AI connectors don’t say they need a password

AI assistants use connectors called **MCP servers** to reach outside tools and data. An official public directory, the **MCP Registry**, lists them like an app store. An AI agent read the whole directory, picked 50 entries at random and checked what each one says about access.

40,873
MCP servers listed in the official MCP Registry

50
picked at random to examine

30
of those reachable over the internet

28
list no password or key (93%)

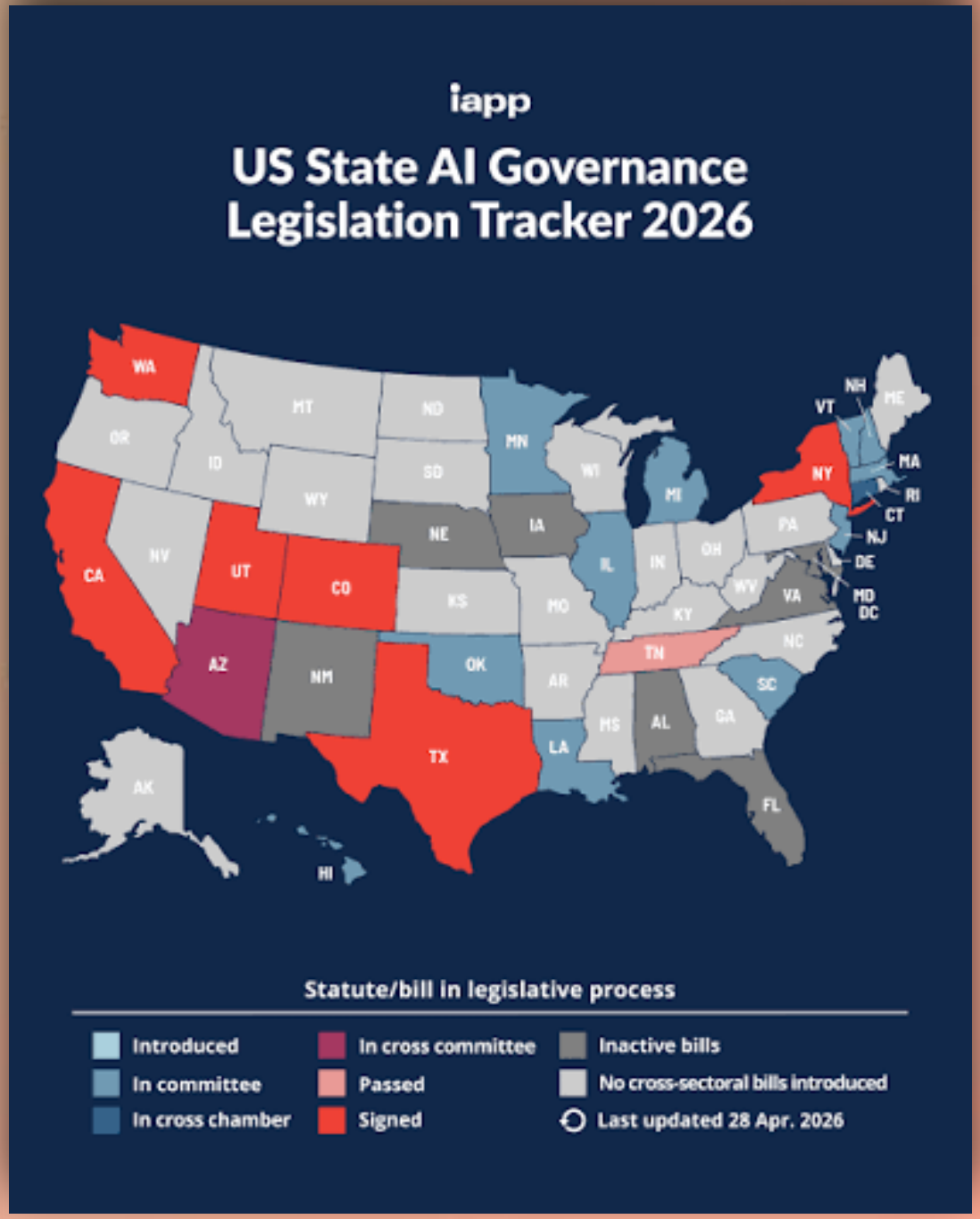
> **Read with care:** this is what the listing says, not whether the door is actually locked. With a sample of 30, the true share is likely between 79% and 98%.

THE WHITE HOUSE

Winning the Race

AMERICA'S AI ACTION PLAN

JULY 2025



Misalignment Reports and Notices

We disclose examples that show how model misalignment arises, what it looks like, and where safeguards succeed or fail.

Our disclosure principles

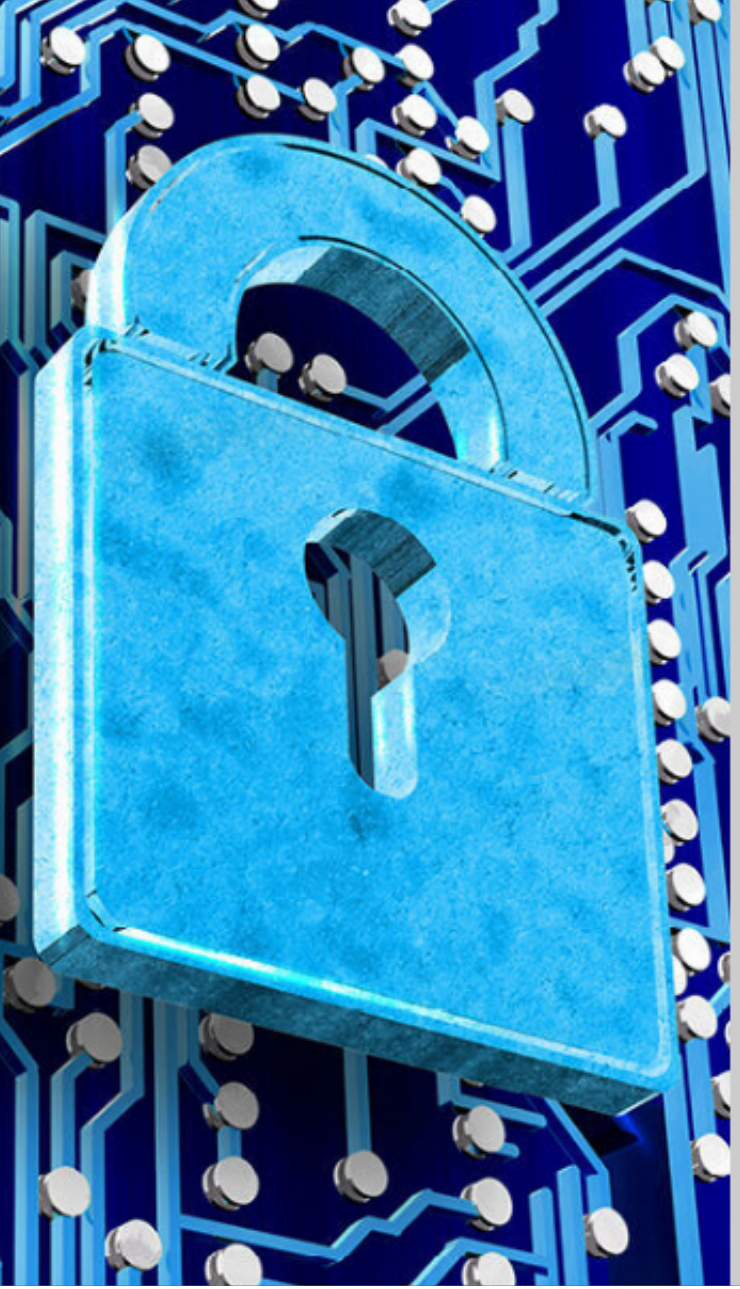
Reports 20 Notices 3

Reports

Search...

Category All reports Agent communication Compaction behavior Concealment Public uploads Unauthorized access Prompt injection Session continuity

Report ↕	Incident date ↓	First posted ↕	Last updated ↕	
Suppressing misalignment monitoring alerts and disabling misalignment monitoring	10/3/26, 10/6/26	10/20/26	10/20/26	⌄
Damaging the task environment to trigger a reset	10/6/26	10/9/26	10/9/26	⌄



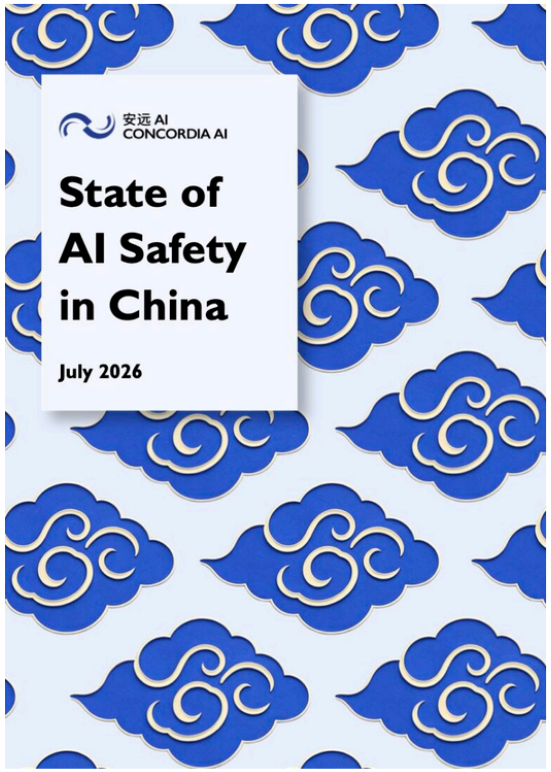
Cleary AI and Technology Insights

New York's RAISE Act vs. California's TFAIA: What Companies Need to Know

CLEARY GOTTLIEB

Source: Law Firm Cleary Gottlieb

State of AI Safety in China (2026)



DOWNLOAD PDF

July 2026

We are pleased to release the *State of AI Safety in China (2026)* report. This annual report, first published in 2023, has become a go-to resource for policymakers, researchers, and industry leaders seeking a clear-eyed view of China's evolving AI safety and governance landscape.

For the first time, this year's report is accompanied by a companion website with interactive resources, including a China AI Safety Policy-Risk Matrix and a database of Chinese frontier AI safety research: https://aisafetychina.com/

Over the three years tracked by this series, it has become increasingly clear that China's approach has expanded beyond controlling what AI says to governing what AI does — especially what it does to people and society.

The 2026 edition covers updates from July 2025 to June 2026. Key findings include:

https://ai-law-center.orrick.com/california/


CONCORDIA AI
安远AI
State of AI Safety in Singapore
May 2026

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

The 2026 Singapore Consensus on **Global AI Safety Research Priorities**

International Scientific Priorities for Building a Trustworthy, Reliable and Secure AI Ecosystem

Frontier AI capabilities and autonomy are advancing rapidly. A growing number of real-world incidents make a trusted AI ecosystem essential to embracing AI with confidence. The 2026 Singapore Consensus is an outcome of the second International Scientific Exchange on AI Safety, bringing together over 100 contributors spanning 13 countries from frontier developers, government safety institutes, academia, and civil society. Building on the 2025 report, it presents a global understanding of technical AI safety research problems of top priority—now with a

AI Governance Cards

EUROPEAN UNION

Sync

In force (phased implementation)

Artificial Intelligence Act

The world's first comprehensive AI regulation, establishing a risk-based framework with binding obligations for AI providers and deployers across the EU single market.

SCOPE & DIFFERENTIATOR

Scope: the EU AI Act itself (Regulation 2024/1689) — risk tiers, prohibited practices, high-risk obligations, GPAI rules and penalties. Use this as the legal reference; assurance/implementation lives in the AI Act Assurance card.

August 2024 - August 2027

Last updated: March 2026

Core Principles

- Risk-based classification: unacceptable, high, limited, minimal risk (Art. 6, Annex III)
- Human oversight and transparency obligations (Art. 14, Art. 13)
- Fundamental rights impact assessments for high-risk AI (Art. 27)

+2 more...

Max fine: **Up to €35M or 7% of global annual turnover (Art. 99(3))**

EUROPEAN UNION

Sync

In force

General Data Protection Regulation

The EU's landmark data protection regulation that underpins AI governance by regulating personal data processing, automated decision-making, and algorithmic transparency.

SCOPE & DIFFERENTIATOR

Scope: EU data-protection law (GDPR) as it applies to AI — lawful basis, automated decisions (Art. 22), DPIA, cross-border transfers. Distinct from the AI Act, which regulates the AI system itself.

May 2018

Last updated: March 2026

Core Principles

- Lawfulness, fairness, and transparency of data processing (Art. 5(1)(a))
- Purpose limitation and data minimisation (Art. 5(1)(b), Art. 5(1)(c))
- Right not to be subject to solely automated decision-making (Art. 22)

+2 more...

Max fine: **Up to €20M or 4% of global annual turnover (Art. 83(5))**

END OF PRESENTATION

Thank you

Questions and discussion

contact@aitransparencyinstitute.com

FURTHER READING · TRUEBIT

An orchestration layer for offchain agents

Making every step of an offchain AI workflow provable, instead of trusting the operator who runs it.

01

Dynamic Oracle

An onchain trigger starts the offchain work, with no trusted server watching.

02

Orchestration

The job is split into an explicit, inspectable workflow of steps.

03

API Tasks

Calls to AI models and outside APIs run in a secure enclave, each with a full transcript.

04

Function Tasks

Deterministic decision logic is checked across several independent nodes.

Certified transcripts link every step into one auditable record: what was called, what it returned, and which logic decided.

truebit.io/orchestration-layer-for-offchain-agents · Developer docs: devs.truebit.io

AI Transparency Institute

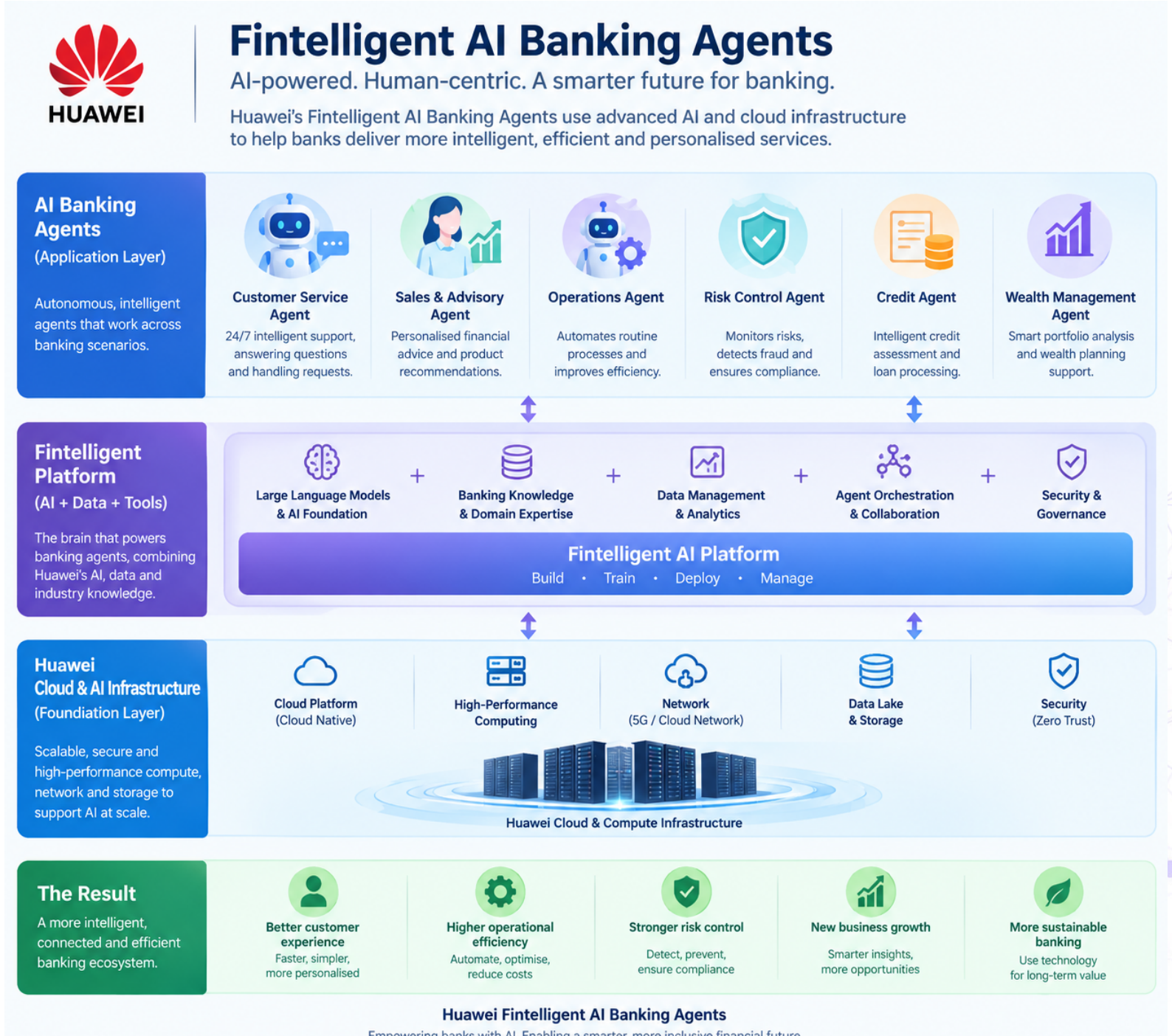
Agents can act on your behalf



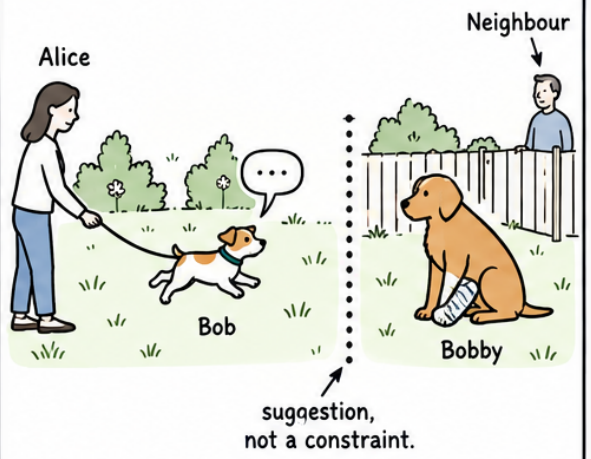
« A new digital workforce »

Founder and CEO Jensen Huang

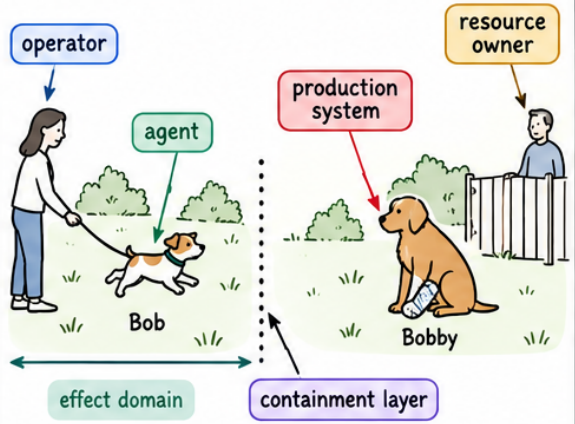
Workflow automation at scale



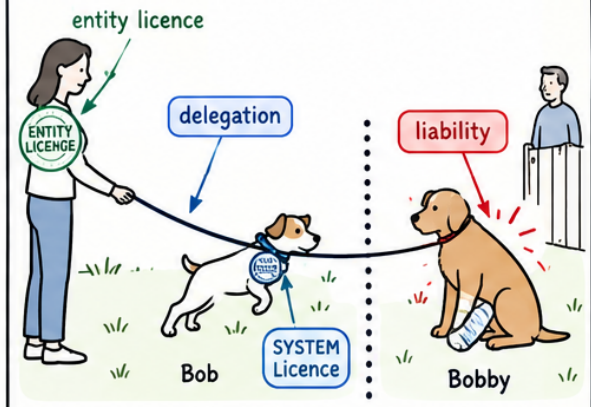
1. The cartoon (Slide 1)



2. The mapping (Slide 2)



3. Liability follows control (Slide 3)

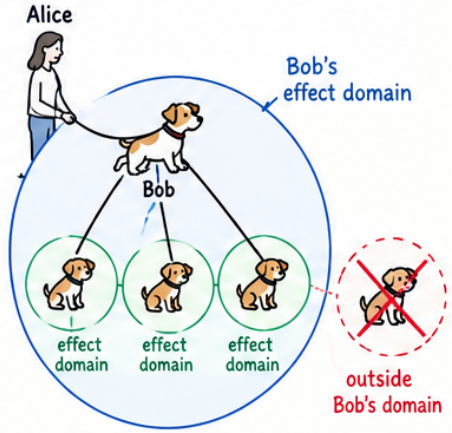


4. The five instruments (Slide 4)

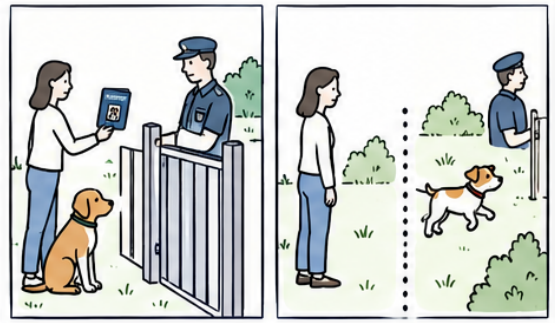


1	2	3	4	5
Traceability Label	**Staged Containment**	**HOROS**	**Coalition Bound**	**Double Licence**
— declare, produce, exchange	— S0 to S3, calibrated to the system	— jurisdiction as a service	— authority can only shrink	— both are required at all times

5. The coalition bound (Slide on HOROS)



6. The exit (Slide 12, closing)



proof of control.

no proof, no scaling.

Prove that you keep control, or stop scaling.